A phonetic Vocoder for Very-Low-Rate Speech Coding

AUTHOR(S):

Nakagawa, Seiichi; Hirata, Yoshimitsu

CITATION:

Nakagawa, Seiichi ...[et al]. A phonetic Vocoder for Very-Low-Rate Speech Coding. 音声科学研究 1989, 23: 44-56

ISSUE DATE:

1989

URL:

http://hdl.handle.net/2433/52492

STUDIA PHONOLOGICA XXIII (1989)

A Phonetic Vocoder for Very-Low-Rate Speech Coding

Seiichi NAKAGAWA and Yoshimitsu HIRATA

ABSTRUCT

In this paper, we describe a phonetic vocoder based on the concatenation of syllable-units which represent speech waves by extremely low rate (100 bits/s) using a speech recognition tequnique. We take syllables into consideration as the unit of recognition/synthesis, because a syllable contains the coarticulation effect between a consonant and a vowel. Speech waves are transformed into a sequence of frames, each of which consists of LPC cepstrum, PARCOR coefficients, pitch and power. After the O(n) DP matching with 500 reference patterns, the input speech is transformed into a sequence of Japanese syllables. The information of recognized syllable contains the category of syllables, duration, power and pitch, and is represented by 16 bits. Using this vocoder, speech can be represented by only 100 bits/sec and the intelligibility of phrase for an unlimited task is about 60%. If the number of references is enlarged, say, 1600 patterns, the intelligibility becomes of more than 70%. In this case, the coding rate is about 112 bits/sec.

1. INTRODUCTION

The extremely low bit rate coding is effective in the reduction of a large quantity of speech data (e. g., voice mail) or mobile communication. For a very low bit rate coding of speech, segment quantization methods based on vector/matrix quantization have been contrived and made good intelligibility at 150 ~ 200 bits/s [1] [2]. But these methods are regarded as a kind of pattern matching vocoders and do not use language knowledges. That's why such a very low bit coding is possible for any language. In order to make furthermore low bit coding possible, we must take linguistic knowledges into coding.

There are some studies on such a vocoder, e. g. a phonetic vocoder which finds a phoneme string to minimize the distance between the input speech and diphone templates and which can represent speech at 100 bits/s in Schwartz et al. system [3]. Because there is a limitation to the next phoneme by a diphone

Seiichi NAKAGAWA (中川聖一) : Associate Professor, Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, 441 Japan.
Yoshimitsu HIRATA (平田好充) : Graduate Student, Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi, 441 Japan.
The authors have done the research under the direction of Dr. Shuji Dohshita, Professor of Information Science, Kyoto University.

network, that is, context-dependency, there is not obtained enough intelligibility. They found that a phonetic vocoder should have a phonetic recognition rate at least 80% and natural phonetic synthesis. The rate of 80% on phoneme recognition corresponds to that of 60 ~ 65% on syllable recognition. It was also reported that the intelligibility of this vocoder was improved by the allowance to follow any templates, that was, a segment vocoder using diphones as segments (200 bits/s) [4]. Another type of phonetic vocoders was proposed recently [5] [6], and they used recognizers based on HMM's to code speech. Picone and Doddington transformed speech (20 frames/s) into a sequence of phonemes (120 bits/s for spectral information) and durations (50 bits/s for state transitions). The phone recognition accuracy was about 35%. The phone pair grammar (phonotactic constraints) did not improve the recognition accuracy and speech quality. This vocoder's quality was comparable with that of a VQ-based vocoder of 300 bits/s. Soong proposed a speech recognition/synthesis method based on 2084 different left and right context-dependent tri-phone model. He obtained the phoneme recognition rate of 96%.

We have two choices for a very low rate speech coding. One is the approach to mininize the spectral distortion. The other is one to improve the recognition accuracy. We chosed former approach.

In this paper, we describe a vocoder which can represent speech at 100 bits/s using a speech recognition technique. The synthetic unit of the vocoder is a syllable. Our study is based on the assumption that 1) human beings has high ability of linguistic understanding for synthesizeg speech in noisy environments even if speech recognition by machine is incomplete, 2) there is no direct relationship between phoneme recognition accuracy by machine and intelligibility by a speech synthesizer. In these assumptions, a segment vocoder may be superior to a phonetic coder. Hewever, it is very difficult to make a codebook of segment patterns with variable lengths. One of most differences is that our approach allows a dynamic time warping for time-series patterns, unlike segment vocoder. (A segment vocoder allows a linear time warping.) First, we describe how speech is coded into an optimal syllable sequence. Next, we report on some experimental results to evaluate this vocoder.

2. Coding Into An Optimal Syllable Sequence

2.1 System organization

Considering that the source of speech in brain is a sequence of discrete symbols, the ultimate coding is obtained by coding into the symbol. There are some units of language such as phoneme, demi-syllable, syllable, word, etc. We consider syllables as a minimum unit of the language which can be dealt with easy in co-ariculation between a consonant and a vowel. The way we use in coding

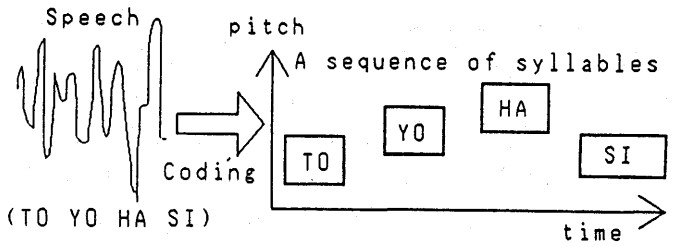

Fig. 1 Coding into an optimal syllable sequence

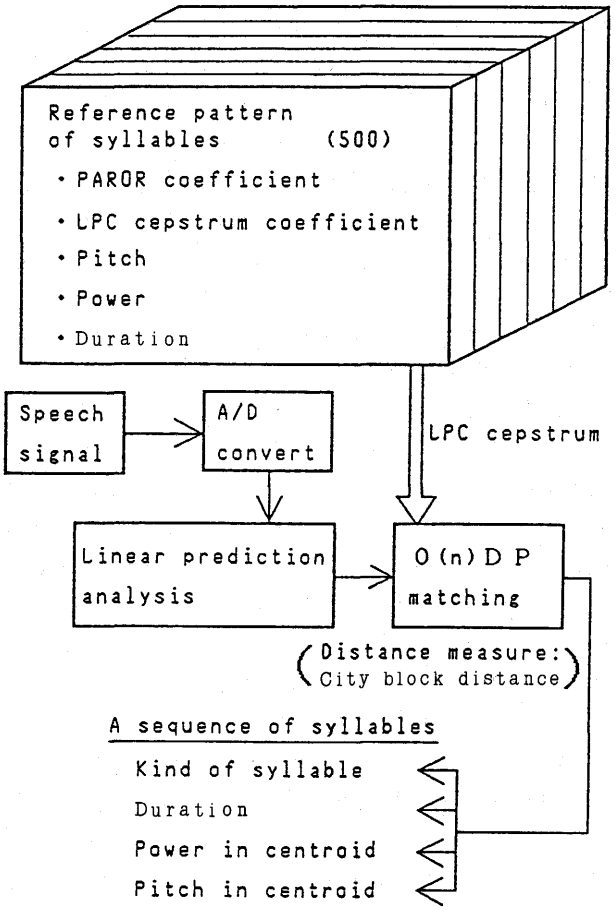

Fig. 2 Block diagram of a phonetic vocoder

speech is shown in Fig.1. Fig. 2 shows a block diagram of the coder.

Input speech waves are sampled at 10kHz in 12 bits with an A/D converter and transformed into a sequence of frames by the 14th oder linear prediction analysis. Each frame has LPC cepstral coefficients, PARCOR coefficients, pitch and power. On syllable recongition, we used the O(n) DP [7] (or one stage DP [8]) matching algorihm which had been used for connected word recognition. After the O(n) DP matching with reference patterns, the input speech is transformed into a sequence of syllables. A recognized syllable contains the category of

syllables, duration, power and pitch. Each item is quantized by the levels of 500, 3, 5 and 7, respectively, that is, the spectral information is quantized by 9 bits/syllable and prosodic information 7 bits/syllable. If the speech rate is 6 syllables/sec, the amount of coded speech is 96 bits/sec.

2.2 Reference pattern

It is known that plural patterns are need in each category as references, because the acoustic property varies in context, in other words, it has allophones. We extracted reference patterns of about 500 syllables (in Japanese, there are about 100 syllables. Almost all syllables consist a preceding consonant and a vowel in Japanese.) from 416 words uttered in isolation. Each syllable corresponds to one CV syllable extracted from all VCV syllables which are included in the 416 words, where C and V denote a consonant and a vowel, respectively. These extraction was performed by hand labeling.

2.3 Coding method by O(n) DP matching method

For continuous speech recognition, we have used the O(n) DP matching method [7]. The O(n) DP matching algorithm is computationaly more efficient than the Two Level DP matching [9] and make an optimal syllables sequence according to a distance measure.

We use LPC cepstral coefficients as a feature parameter, it is better than PARCOR coefficients from recognition accuracy view. And we use the city block distance (Chebyshev norm) as a spectral distortion measure.

2.4 Extraction and coding of prosody

As we described above, after the recognition with the O(n) DP matching, input speech is transformed into a sequence of syllables which have category of syllables and prosodic information. In this section we describe how to extract and code the information of prosody which consists of pitch, power and duration.

Pitch is represented as one in the centroid of syllable part, and the value divided by the pitch in the preceding syllable is quantized by 7 levels as follows:

	~ 100/125	→	100/130
100/125	~ 100/115	→	100/120
100/115	~ 100/105	→	100/110
100/105	~ 100/ 95	→	100/100
100/ 95	~ 100/ 85	→	100/ 90
100/ 85	~ 100/ 75	→	100/ 80
100/ 75	~	→	100/ 70

A pitch of centroid is defined as an average of a section from 1/2 to 5/6 of syllable obtained by the recognizer. In this way we need some representation of the first syllable pitch, so it is represented as the ratio to a constant/standard value.

Power is quantized as the ratio of the power in the centroid to the corresponding one of the reference pattern by 5 levels as follows:

	~ 0.3	→	0.1
0.3	~ 0.75	→	0.5
0.75	~ 3.0	→	1.0
3.0	~ 7.5	→	5.0
7.5	~	→	10.0

There is another way to represent the power, that is, the absolute power value is quantized. If we use the latter, we need many of levels to quantize the power, so extremely low bit coding doesn't come true. Assuming the ratio of power to one of the corresponding reference is in small range, it is better to use the former. Because there is a high correlation between power and pitch, it is enough to quantize power at 3 levels. However we have not yet used the correlation.

Duration is represented as the ratio the length of the syllable obtained by matching to that of the corresponding reference pattern, and the ratio is quantized by 3 levels as follows:

~ 6/6	→	5/6
6/6 ~ 9/6	→	8/6
9/6 ~	→	11/6

After each syllable can be represented by 16 bits, and the speech which duration is one second can be represented by only 96 bits if speech is spoken at a speed of 6 syllables/s.

3. DECODING METHOD

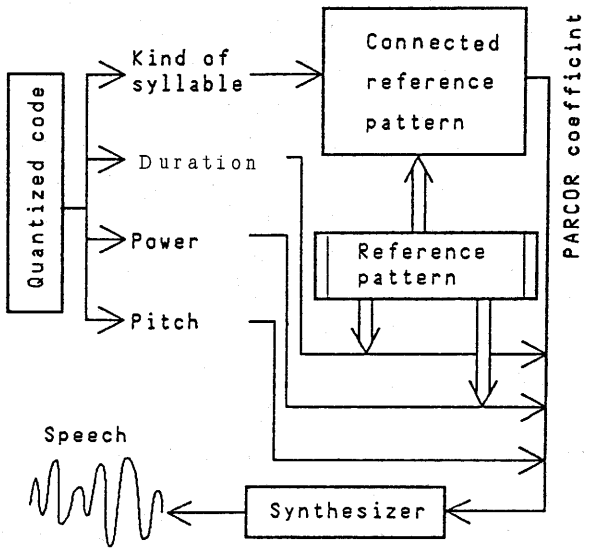

Fig. 3 Decoding method

Fig. 3 shows the block diagram of decoder. For a sequence of quantized codes which are obtained by a recognizer, speech is decoded by concatenating the

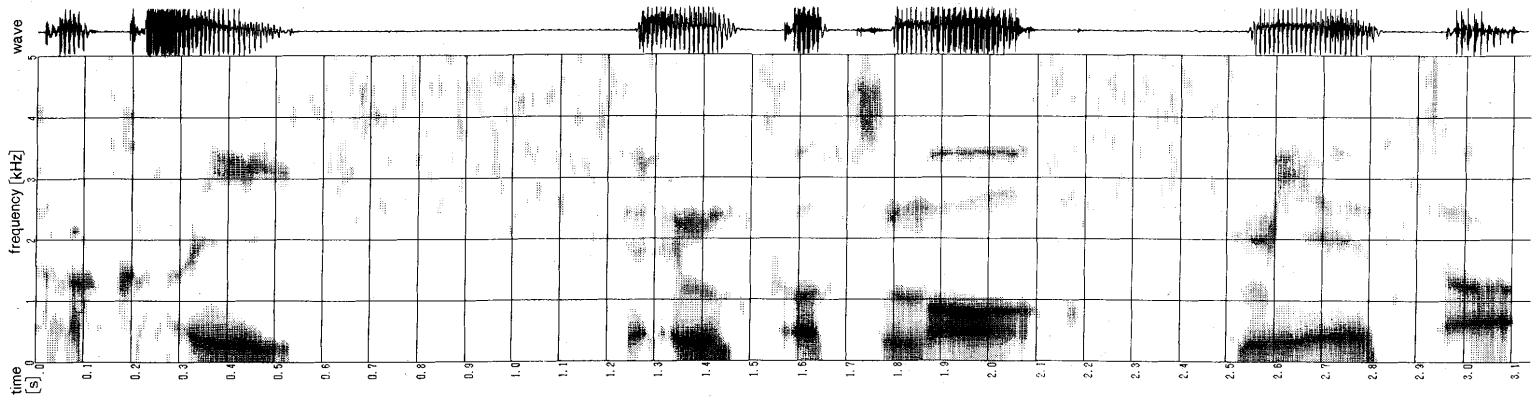

Fig. 4 (a) Sonagraph of Input (Original Speech : takai eNtotsu-mo mieta)

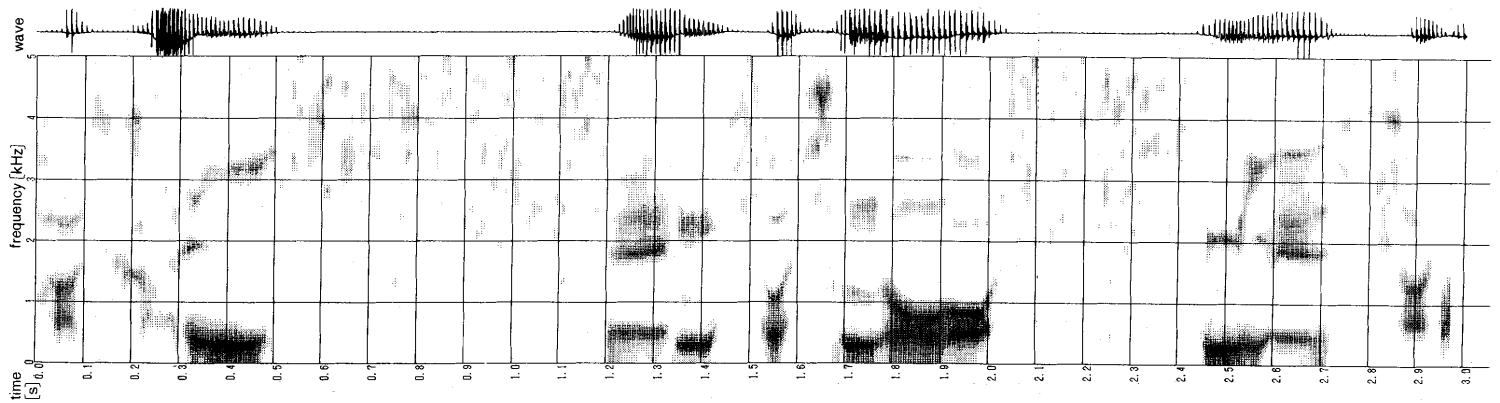

Fig. 4 (b) Sonagraph of Synthesized Speech (Recognized Syllable Sequence : ha ka i e N to - no o mi e sa a)

corresponding reference patterns. We use a CV-concatenation method essentialy and linear-interpolate PARCOR coefficients between syllables. Let the syllable boundary be a point between (n−1)th frame and n-th frame, and [n] be PARCOR coefficient and power. The interpolation is given as follows:

$$[n-2] = [n-2] \times 0.7 + [n-1] \times 0.2 + [n] \times 0.1$$
$$[n-1] = [n-1] \times 0.6 + [n] \times 0.3 + [n+1] \times 0.1$$
$$[n] = [n-2] \times 0.1 + [n-1] \times 0.2 + [n] \times 0.7$$

Because each CV syllable has several reference patterns, the reference patterns corresponding to recognition result are adopted. Assuming that input speech is "TO YO HA SI", for example, a desired recognition result may be "TO $_{A}$YO $_{U}$TA $_{A}$SI" ("$_{A}$YO" means a CV syllable "YO" which is a part of VCV triphone "AYO"). Considering the result is optimal from the view to minimize spectral distortion, it is unnecessary to recognize as "TO $_{O}$YO $_{O}$HA $_{A}$SI". This corresponds to the fact that the phonotactic constraints do not improve the speech quality or intelligibility [4, 5]. And it is unnecessary to correct a syllable sequence because the decoded speech has high ability to be understood as the origin by a human perceptibility if the spectral distortion is remarkably small.

Duration in decoding is modified by eliminating or repeating a part of the vowel (a section from 1/2 to 5/6 of a syllable) of the reference pattern, for it may be hear as another phone if a consonant part is warped.

Fig. 4 illustrates an example of coding and decoding. The input sentence is (No. 19) "takai eNtotsu-mo mieta (We also saw a high chimney)". The transformed (recognized) syllable sequence obtained by using 1634 patterns is "ha ka i e N to no o mi e sa a." Fig. 4(a) shows the sonagraph of original speech. On the other hand, Fig. 4(b) shows the sonagraph of synthesized speech.

4. EXPERIMENTS

In a preliminary experiment, we tested the intelligibility of this vocoder for limited 100 spoken words. The subjects identified the presented word at correct rate of about 96% when it was tought in advance that the word was one of the list of 100 words [10]. In this section we describe several expriments in order to test the feasibility for sentences of the vocoder.

First, the intelligibility of synthetic voice was compared with the case of a VQ-based vocoder and without segmentation errors in the syllable recognition. Secondly, the pitch information of speech was added in decoding. Third, we enlarged the number of syllable reference patterns. Finally, we tried to convert voices to a standard speaker's one who was different from a parson in coding.

4. 1 In the case of perfect segmentation

In order to assess an upper-bound ability of this vocoder, we checked the intelligibility of the vocoder in the no case of segmentation error. The O(n) DP

algorithm was modified not to make any syllable boundary detection except them given by manual segmentation. We used some sentences spoken with pauses between phrases at a speed of 6 ~ 7 morae (12 ~ 14 phonemes) per second. In this experiment, we used about 85 bits/s coding rate except for the pitch information.

Fifteen sentences spoken by a male speaker HN who uttered isolated words for reference patterns were recognized in two ways. One was the O(n) DP matching and the other was the matching with given boundaries. The recognition result is shown in Table 1.

Table. 1 Syllable Recognition Results
(Speaker : HN)

method	O(n) DP	boundary known
syllable recognition rate	53%	69%
insertion error rate	25%	0%
deletion error rate	2%	0%
segmenation rate	73%	100%
coding distortion	1.46	1.65

The coding distortion means an average distance normalized by the input frame length on matching according to city block distance. The segmentation rate is defined as follows :

$$\text{Segmentation rate} = \frac{\text{INP} - (\text{INS} + \text{DEL})}{\text{INP}}$$

INP : Number of input syllables
INS : Number of inserted syllables
DEL : Number of deleted syllables

By giving boundaries, the syllable recognition rate was improved from 53% to 69%. The reason why the coding distortion of the O(n) DP matching is smaller than one of boundary known case is that the O(n) DP matching algorithm concatenates the best reference patterns and marks boundaries in input speech to minimize the distance between input speech and any string of references.

Table 2 shows which level of pattern matching vocoder based on vector quantization corresponds to or is equivalent to the distortion of the O(n) DP matching, where it is obtained from three sentences "Hana yori dango" ("Cake is prefered to flower" in English), "Migini sanjudo maware" ("Rotate to right by 30 degrees") and "Itokowa sizukana ongakuga totemo sukidesita" ("My first cousin liked soft music").

We took place a hearing test in the way as follows : The synthesized voices for fifteen sentences were presented by eight subjects. Each of the sentences is simple but the task is not restricted. Five sentences among the fifteen sentences were synthesized by the syllable vocoder with unknown boundaries (pitch is constant),

Table. 2 Coding Distortion
(Distance measure : city block distance of PARCOR coefficients)

codebook size	distortion	spectral information
16 (VQ)	1.52	400 bits
32 (VQ)	1.39	500 bits
64 (VQ)	1.29	600 bits
128 (VQ)	1.20	700 bits
256 (VQ)	1.14	800 bits
O(n) DP	1.54	54 bits

the same vocoder with known boundaries and a pattern matching vocoder based on vector quantization (codebook size of 64, 600 bits/s), respectively. The subjects who don't know the contents of sentences listened twice per one synthetic sentence, then they dictated sentences at their listening. As results, the phrase inteligibility was about 60% for the three vocoders.

The coding distortion of the O(n) DP matching corresponds to one the pattern matching based on vector quatization in 16 codebook size. But the intelligibility of the O(n) DP matching vocoder was same as a pattern matching vocoder quantized at 64 levels. In the no-error segmentation, the intelligibility was the same as unknown boundary case in spite of our expectation that the intelligibillity would be improved by rising the recognition result. The reason is considered as follows : even if the recognition accuracy is not good there may be a little influence on human perception because a spectral sequence obtained from the O(n) DP matching preserves the linguistic information, in particular, the dynamic information. In other words, the listener prefers smaller spectral distortions to higher syllable recognition rate. This advantage is the same as that of a segment vocoder.

4. 2 Addition of pitch information

It is important to use prosodic information to improve the syllable recognition rate or to bring on the naturality. We compared the intelligibility of the vocder by a constant pitch with one of the vocoder by the pseudo-real contour pitch. The pitch period was given for three points per one recognized syllable. Each of them was quantized at 7 levels in the same way described before. In decoding, we linear-interpolated pitches at the point between them.

We prepared 28 sentences and reference patterns spoken by each of male speaker YH and female speaker TM, respectively. The inteligibility of O(n) DP matching vocoder was tested for both cases of the constant pitch and variable pitch. To compare them the same condition was hold for a pattern matching vocoder based on the vector quantization method (code book size is 64 and the coding rate is 600 bits/s. When the pitch information is added to every frame, the

coding rate is about 1200 bit/s.). For sentences spoken by speakers TM and HN, the number of subjects was eight, for speaker YH, five. Table 3 shows exprimental results. We should notice that the stimulated sentences for each item in experiments were different from one another. Therefore, the intelligibility depends more or less on the contents of sentences. For both male speaker YH and female speaker TM, the intelligibility of the O(n) DP matching vocoder was improved by adding pitch information. In particular, the intelligibility of VQ vocoder was remarkably improved. It was caused by that the PARCOR coefficients (spectral information) were influenced by the vocal source (pitch or prosidic information), in other words, there was a correlation between PARCOR coefficients and pitch. But there was no change for speaker HN. Besides the syllable recognition result was about 51% recognition rate, 87% segmentation rate for 28 sentences spoken both speakers YH and TM.

Table.3 Phrase Intelligibility

type	pitch	speaker		
		YH	TM	HN
O(n) DP	constant	55%	53%	59%
	variable	67%	56%	60%
VQ	constant	80%	74%	60%
	variable	99%	93%	—

4. 3 Enlargement of reference patterns

In this section, we describe the experimental result by the enlargement of refference patterns. The syllable reference patterns were extracted from 416 words uttered in isolation. These words contain 1634 syllables in total. We adopted these all syllables as reference patterns. The detailed results of decoding are shown in Table 4. The average recognition rate and segmentation was improved to 59.2% and 83.6% from 53% and 73%, respectively (see Table 1). The average distortion corresponds to the vector quantization in 32 codebook size.

The phrase intelligibility is shown in Table 5. The number of subjects was five described obove and three new subjects (YU, TT, TH) were unfamiliar with synthesized speech. The intelligibility remarkably improved to move than 70% by enlarging the number of refference patterns. In this case, the coding rate is 112 bit/s.

4. 4 Voice conversion to a standard speaker's voice

To put such an extremely low bit rate coding into practice, the decoding speech using a speaker's voice who is same as one in coding is not a good way because of neccesary of transmitting reference patterns. This problem can be avoided by preparing one set of a standard speaker's reference patterns beforehand to concatenate them according to a sequence of syllables which have kinds of syllables and prosodic information.

Table. 4 Syllable Recognition Results by O(n) DP Matching with 1634 Reference Patterns (Speaker : HN)

sentence number	number of syllable	correct	insertion	deletion	segmentation rate	syllable recognition rate	distortion
2	22	14	2	0	90.9	63.6	1.51
5	24	13	1	2	87.5	54.2	1.48
10	21	16	1	0	95.2	76.2	1.33
11	24	12	3	2	79.2	50.0	1.48
13	21	11	1	2	85.7	52.4	1.37
14	28	15	4	0	85.7	53.6	1.35
15	20	14	1	1	90.0	70.0	1.37
17	18	11	2	0	88.9	61.1	1.36
19	11	7	2	1	72.7	63.6	1.35
20	19	13	4	0	78.9	68.4	1.46
21	17	11	3	1	76.5	64.7	1.37
22	19	10	2	0	89.5	52.6	1.39
23	15	6	2	0	86.7	40.0	1.30
27	16	12	5	0	68.8	75.0	1.30
28	17	8	5	1	64.7	47.1	1.30
average (%)			13.1	3.4	83.6	59.2	1.39

Table. 5 Phrase Intelligibility (number of reference patterns=1634)

method	sentence No.	number of phrase	number of correct phrases every subject						average (%)
			YU	MT	SY	TT	IH	TH	
O(n) DP pitch= constant	2	5	3	0	4	4	5	2	60.0
	5	6	5	6	6	6	6	5	94.4
	11	5	1	2	5	5	5	2	66.7
	14	6	5	5	6	6	5	6	91.7
	15	5	5	4	5	4	5	5	93.3
	19	3	0	1	1	1	2	1	33.3
	20	4	4	4	4	4	4	4	100.0
	22	5	2	4	1	1	3	1	40.0
average (%)			64	67	82	79	90	67	74.8
O(n) DP pitch= variable	10	5	4	5	5	5	5	4	93.3
	13	6	6	5	6	6	6	2	86.1
	17	5	5	5	5	3	3	3	80.0
	21	5	2	3	1	2	4	3	50.0
	23	4	2	3	4	4	4	3	83.3
	27	6	6	6	6	5	6	6	97.2
	28	5	3	3	2	3	3	2	53.3
average (%)			78	83	81	78	86	64	78.2

Fig. 5 shows how to decode as a standard speaker's voice. Using recognition results, the reference patterns of the speaker in coding are replaced with those of the standard speaker. Input speech is coded by the method stated previously. On side of receiver, the standard speaker's PARCOR coefficients are concatenated coresponding to kinds of syllables. Pitch is multiplied by a ratio of the average standard speaker's pitch to the average input speaker's pitch.

We tested the intelligibility in the case of converting speaker HN's voice into speaker YH's voice. Five subjects listened to synthetic speech (pitch is constant) twice and dictated what was said. The phrase intelligibility was 46%. Since the original speaker HN's one was about 60% as shown in Table 3 with no pitch information, there was degradation by converting into another speaker's voice. A reason for this result is considered as follows : there is no problem in listening for original synthesized speech because the matching is done in order to minimaize distances between input and reference patterns, but the converted spectral sequence may not be optimal in replacing into other speaker's reference patterns.

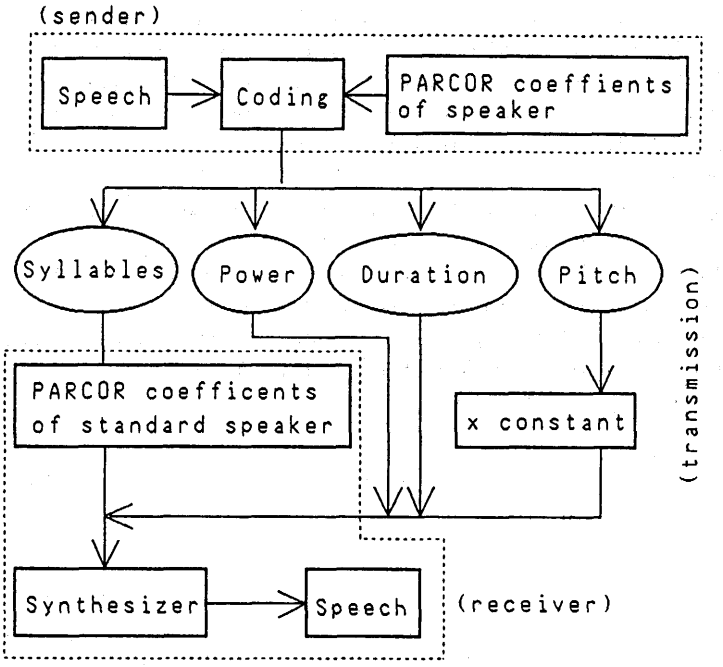

Fig. 5 Synthesis as a standard speaker's voice

5. Conclusion

We described how to realize a 100 bit/s coding of speech. The purpose of this vocoder is a study of representing speech as a text file. Finally the phrase intelligiblity of this vocoder was a little less than 70% in the coding rate of 100bit/s and a little more than 70% in the 112 bit/s. In order to imporove the

intelligibility, following points are considered as effective : enlargement of refference patterns, introduction HMM's to imporove recognition results [11] or to extract automatically reference patterns [12], and representing speaker characteristics in low bit for unspecified speakers.

REFERENCES

[1] S. Roucos, R. M. Schwartz and J. Makhoul : "A Segment Vocoder at 150 B/S", Proc. ICASSP, pp. 61-64 (1983).

[2] Y. Shiraki and M.Honda : "LPC Speech Coding based on Variable-Length Segment Quantization", IEEE Trans. Acoust. Speech & Signal process., ASSP-36, 9, pp. 1437-1444 (1988).

[3] R. Schwartz, J. Klovstad, J. Makhoul and J. Sorensen : "A Preliminary Design of a Phonetic Vocoder based on a Diphone Model", Proc. ICASSP, pp. 32-35 (1980).

[4] S. Roucos, R. Schwartz and J. Makhoul : "Segment Quantization for Very-Low-Rate Speech Coding", Proc. ICASSP, pp. 1565-1568 (1982).

[5] J. Picone and G. R. Doddington : "A Phonetic Vocoder", Proc. ICASSP, pp. 580-583 (1989).

[6] F. K. Soong : "A Phonetically Labeled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis", Proc. ICASSP, pp. 584-587 (1989).

[7] S. Nakagawa : "Connected Spoken Word Recognition Algorithms by Constant Time Delay DP, O(n) DP and Augmented Continuous DP Matching", Information Sciences 33, pp. 63-85 (1984).

[8] H. Ney : "The use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans. Acoust., Speech & Signal Process., ASSP-32, 2, pp. 263-271 (1984)

[9] H. Sakoe : "Two-level DP-Matching-a Dynamic Progrmming based Pattern Matching Algorithm for Connected Word Recognition", IEEE Trans. Acoust., Speech & Signal Process., ASSP-27, 6, pp. 588-595 (1979).

[10] S. Nakagawa and T. Yasumoto : "A Speech Recognition Vocoder (100 bit/s)", Proc. 1988 Spring National Convention IEICE, SA-4-9 (in Japanese).

[11] S. Nakagawa and Y. Hashimoto : "A Method for Continuous Speech Segmentation using HMM", Proc. Int. Conf. Pattern Recognition, pp. 960-962 (1988).

[12] S. Nakagawa and H. Nakanishi : "Speaker-Independent English Consonant and Japanese Word Recognition by a Stochastic Dynamic Time Warping Method", Trans. Inst. Elect. Tel. Comm. Engrs., Vol. 34, No. 1, pp. 87-95 (1988).